Documenting two histories at once: Digging into archaeology

Jon Holmen
The Unit for digital documentation, The Faculty of Art, University of Oslo
Christian-Emil Ore
The Unit for digital documentation, The Faculty of Art, University of Oslo
Øyvind Eide
The Unit for digital documentation, The Faculty of Art, University of Oslo

In this paper we will give an overview of the ways in which the principles of edition philology have been applied to archaeological texts and archives. We will also discuss what way we may encode the semantic content of a text in order to make it possible to get useful answers to queries as well as to map the information into a formal standard ontology (the CIDOC-CRM). We have used SGML to express the semantic as well as the structural information of archaeological texts.

Introduction

In 1992 the National Documentation Project was launched. A major part of the project was to create an information system for the archaeological museums in Norway. For almost 170 years the archaeological museums in Norway have published information on a yearly basis on their acquired artefacts in specially prepared acquisition catalogues. The descriptions of finds in these catalogues are quite longwinded, including extensive information on the finds, the find contexts, their place and time, the finder or excavator, as well as detailed descriptions and classifications. The series of catalogues served for practical purposes as the main artefact inventory of each museum.

Over the years, computer applications in archaeology have primarily been used to generate a wide variety of statistics as well as to develop inventory databases. Pattern recognition and simulation/AI were at their peaks around 1980 and 1990 respectively. In the 1990s GIS, 3D modelling and the Web were in focus (see Scollar 1999). Most applications were designed to analyse information collected during fieldwork. Archaeologists use texts in the same way as other scholars, but text philology as such is not central central part of their discipline. Thus the task of creating a database on the basis of old reports is normally done through reading the text and keying into a database the information considered to be essential in a normalised form.

The Documentation Project covered several disciplines including electronic text collection. Thus it was quite natural to try to apply principles from the text encoding community and to use SGML to mark-up the catalogue texts. In 1992 the text encoding community had little to offer to archaeology. Archaeological reports seem not to be the focus of interest for text scholars. As a consequence we had to write our own DTDs covering the classification practice over the past 170 years. In 1992-2000

almost 30,000 printed pages of text were converted and SGML tagged (see Holmen and Uleberg 1996, Holmen and Ore 1996 and Ore 1998 for more details on the project).

```
<CATYEAR>
<INTRO> <MNAME>UNIVERSITETETS OLDSAKSAMLING TILVEKST
</MNAME><YEAR>1989 </YEAR></INTRO>

....
<NRPAR><CATNR nrid="37267">C.37267.
<ARTEFDATA><ARTEFACT>Axe</ARTEFACT> of <MAT>iron</MAT>
<SHARED>from <PERIODE>Late Medieval time</PERIODE>.
<ARTEFDATA><MEAS>L: 141mm</MEAS>, <MEAS>edge W:109mm</MEAS>.
Carpenter's axe with <FORM>specially shaped blade to accommodate the fingers
</FORM> when it is held just "behind" the edge. Particularly necessary for fine work,
when used at an angle to the edge, or as a gouge. <SHARED>Found<FINDLOC>on
the hill, about 300m above <LOC>ÅROS KAPELL</LOC> </FINDLOC>,
<FARM>SJØGLØTT</FARM>, <PARISHD><PARISH>ÅROS </PARISH>
</PARISHD>, <MUNICIPALITY>RØYKEN</MUNICIPALITY>,
<COUNTY>BUSKERUD</COUNTY></FINDLOC>, in <FINDYEAR>1959
</FINDYEAR> by <FINDER>Berge Narvik</FINDER>, Tjernsrudveien 24, Jar,
Oslo. </SHARED></NRPAR>
```

Figure 1 - The SGML mark-up of an acquisition catalogue entry (translated into English)

Over the last couple of years there has been an increasing awareness of the need to include the information and content found in older archaeological and cultural historical oriented documents into archaeological and cultural heritage systems. Our method of encoding and extracting information from electronic versions of old archaeological reports has been taken up by others (e.g. Crescioli, D'Andrea and Niccolucci 2002). Schloen 2001 suggests an XML formalism for storing and interchanging archaeological information. Meckseper 2001 describes the situation in the UK and points out the usefulness of XML. Both seem to address the question of how to use XML for writing and storing new archaeological documents.

A methodological note: Scholarly text editions

In traditional text philology the aim is to create a "best text" based on a set of manuscripts (text witnesses). A scholarly text edition is usually a printed version of the "best text" accompanied by a critical note apparatus documenting variation in the text witnesses and possibly a set of facsimiles of (some parts of) the manuscripts. During the last 10-20 years computers have been introduced into text philology. Today a modern electronic scholarly text edition does not necessarily contain a "best text". A "text edition" consists of a bibliographical database, electronic facsimiles and transcripts of the

text witnesses combined with a few extra search tools and explanatory texts. The transcripts are usually given a mark-up (e.g. XML) that enables hyper-linking as well as a presentation of different views on the texts.

To the extent that text scholars are interested in texts formed at museums, this does usually not extend beyond texts that are first class museum objects. They do not so often consider newer texts describing the normal activity of the museum such as archives, excavation reports etc. to be of interest. We have chosen a wider approach. The internal archives, reports, acquisition catalogues and other documents are the main sources in the development of scientific/scholarly databases for students, scholars and curators at a museum. As a result of our work during the last ten years, we have refined upon a four-step work procedure for the museum texts developed from and inspired by our close contact with other text based disciplines in the humanities. The first three steps correspond to the steps normally applied to an ordinary text edition. The final step, in which we interpret the semantic content of the text according to an ontology, is not so commonly used by text editors.

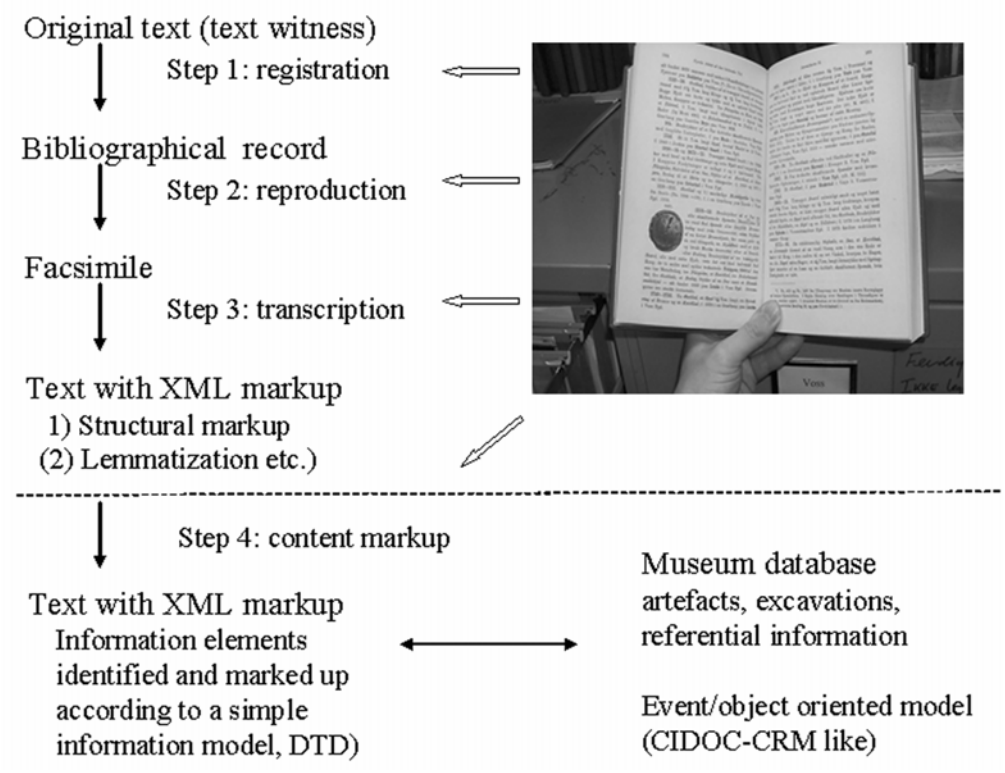

Figure 2 - Digitization of museums documents - ideal versus reality

Methods of digitization

The methodology described in this paper covers the digitization of three main types of archaeological: Acquisition catalogues, a series of printed site and monument records (SMR) and the so-called topographical archives. As the latter of these have not been published as printed books, but comprises large heterogeneous document collections, it was simply too expensive to convert them into electronic

text. These archives were digitised as facsimiles and indexed according to a predefined data model (step 1, 2 and 4 below), see Holmen and Innselset 2003. The acquisition catalogues and the SMR, on the other hand, were OCR read and SGML tagged.

The digitization was performed according to the following four steps:

1 We created a bibliographic record of the original documents including metadata such as Title, Author, Year of publication, Edition number, etc.

2 According to good practise one should make an electronic facsimile by digitising an image of every single page of documents. Unfortunately this was not done due to high storage costs in the early 1990s.

3 We then created an SGML text by applying OCR to the facsimiles or, in the case of handwriting, by manually transcribing the text. The transcript was then given a mark-up indicating pages, special layout, typefaces, illustrations, inserts etc.

4 Finally, we carried out a step which extended the basic descriptive encoding, as we introduced a semantic encoding of the content. This meant that we looked at the information elements in the text and coded them in relation to a given set of concepts or ontology. In our case an ontology is the set of entities and relations that we consider to would best describe the archaeological finds and objects.

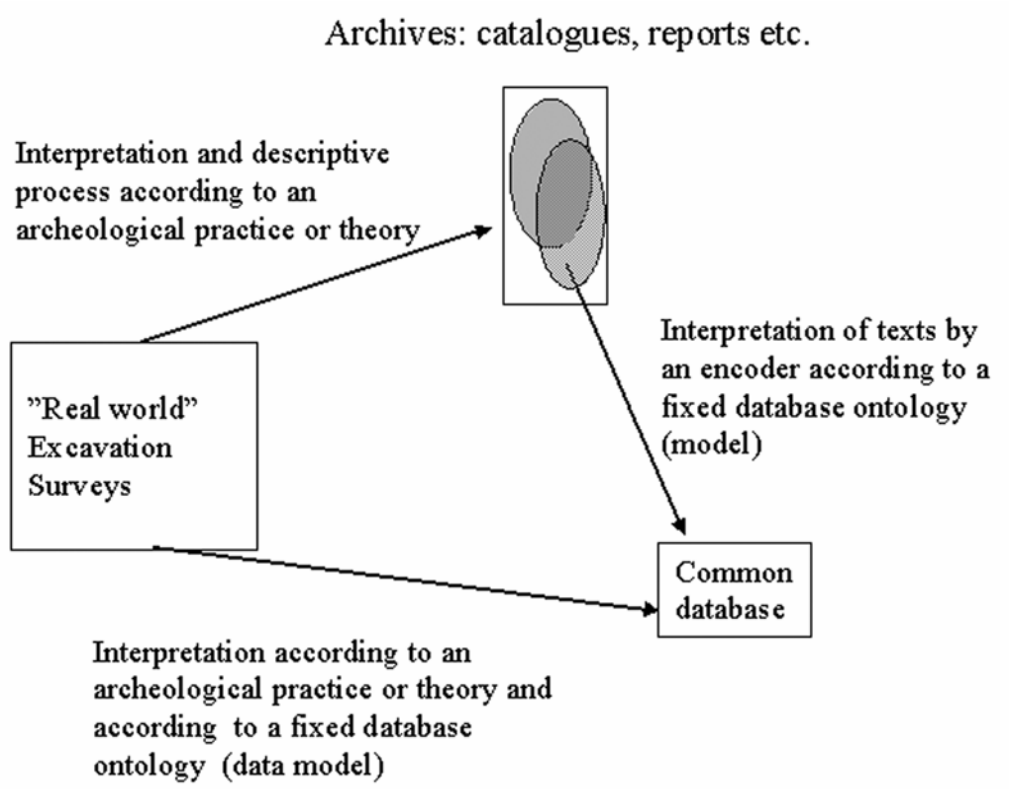

Figure 3 - Information extraction – a multiple interpretation problem

Ontology

The term ‘Ontology’ is a philosophical term denoting the study of being. The Artificial Intelligence community tends to let this term denote a specification of conceptualizations or simply a model of some part of the reality. In computer science the term 'ontology' is often synonymous with 'data model'.

The archive material accumulated during the last 180 years illustrates several ontological shifts in the purely philosophical meaning of the word ‘ontology’. The scholars' own view on the present world and their implied view on the near and distant past have not been constant over the years. As a result the reading of old reports may change. This is a well known fact and one that is usually taken care of by the note apparatus in scientific papers. However, in most databases one usually does not state or comment on the source of each information unit. A not uncommon complaint among non-computerized scholars has been “that information looks so true when seen on a computer screen (but this is not the case)”. Even in the CIDOC-CRM model reification of data is not a part of the model.

The methods found in scholarly text editions and the mark-up process described above are used to trace the source and the interpretation of the original information sources. The encoding of the semantic content must necessarily reflect our view on what constitutes important archaeological information and does not necessarily coincide with the conception of the 19th century author of the original material. The original author included information according to his conceptions, from which we encode the subset that match our conception or ontology. Thus it is important to include at least facsimiles of the original text in order to give the future reader a chance to judge for him- or herself.

As the old material was produced over a period of more than 180 years, it varies significantly with respect to the type of information that is emphasized as well as to style. The methodology used in converting the material gives us additional information. Registers of persons who played an important role in the history of the museums have been produced on the basis of the mark-up. We can also create sequences of the events in the museums’ history, thus giving us the opportunity to realize to what extent each individual's "style" influences the content. By connecting all this we can reconstruct parts of the history of the institutions.

The information about new material and new activities at the museums is treated according to a much more unified standard. This standard is necessarily based on the present archaeological school(s) and will probably be changed in the future when this is considered necessary. Thus we are developing a system in which standards are represented as form definitions/DTDs and preserved.

New material included at a later stage will thus conform with one specific standard that is stored in the database, together with information such as author and date. One implication of this is that when

upgrading the standard, there will be no need for upgrading older material that was entered according to an old standard. We will only need to perform a mapping of the old and the new standard.

This will ensure the preservation of data in its original form, and assure that no information is lost during upgrades. It will also give us and future users of the systems an overview of the development of archaeology as a science and reveal according to which school the various reports were written.

Conclusion

Through the process of converting the original paper-based catalogues and archives at the archaeological museums into digital form, and as far as possible observing our defined principles for this kind of conversion, we have created reliable relational databases of archaeological archives and artefact information. We have also developed an opportunity of working directly with the source material by making an electronic text archive with powerful indexes for both the catalogues and the documents. This makes it possible to study them as individual historical objects.

Through the coding of the source material in a comprehensive ontology we have also, in addition to providing "normal" database functionality, laid the ground for enabling us to describe the history of the museums and the history of Norwegian archaeology. We have also established a system in which the principles used by archaeologists in their documentation work of today may be continually recorded, thus creating what in the future will be history.

References

CRESCIOLI, M., D'ANDREA, A. and NICCOLUCCI, F., 2002. XML Encoding of Archaeological Unstructured Data. In: *Archaeological Informatics: Pushing the Envelope, Proceedings of CAA 2001,* Oxford:267-275.

CIDOC-CRM. *CIDOC Conceptual Reference Model*. Proposed ISO 21127 (http://cidoc.ics.forth.gr/).

HOLMEN, J. and ULEBERG, E. Getting the most out of it - SGML-encoding of archaeological texts. *Paper at the IAAC'96,* Iasi, Romania (http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html).

HOLMEN, J. and ORE, C.-E., 1996. New life for old reports - The Archaeological Part of the National Documentation Project of Norway. *ALLC-ACH '96, Book of Abstracts: Conference abstracts, posters and demonstration*, No. 70, *Report Series of the Norwegian Computing Centre*, Bergen.

INNSELSET, S. and HOLMEN, J., in print. Digital Archaeological Resources at the University of Bergen: An Efficient Tool in Research and Heritage Management? *Paper at the IAAC'99,* Dublin, Ireland.

MECKSEPER, C., 2001. XML and the publication of archaeological field reports. Master's dissertation, University of Sheffield.

ORE, C.-E., 1998. Making multidisciplinary resources. In Burnard, L., Deegan, M. and Short, H. (eds.), *The Digital Demotic, A selection of papers from Digital Resources in the Humanities 1997*, Publication 10, Office for Humanities Communication, King's College, London.

SCHLOEN, D., 2001. Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities* 35:123-152.

SCOLLAR, I., 1999. Twenty five years of Computer Applicatins to Archaeology. In *Archaeology in the Age of the Internet*, BAR Int. Series 750.